LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-681108

Technical Report: Benchmarking for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data

K. McLoughlin

January 22, 2016

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-681108

Technical Report: Benchmarking for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data

DHS Bioforensics Program

IAA No.: HSHQPM-13-X-00219

Principal Investigator and Correspondent

Kevin McLoughlin

Lawrence Livermore National Laboratory (LLNL), Livermore, CA

925-423-5486, mcloughlin2@llnl.gov

Submission date: January 20, 2016

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

1 Introduction

The software application "MetaQuant" was developed by our group at Lawrence Livermore National Laboratory (LLNL). It is designed to profile microbial populations in a sample using data from whole-genome shotgun (WGS) metagenomic DNA sequencing. Several other metagenomic profiling applications have been described in the literature. We ran a series of benchmark tests to compare the performance of MetaQuant against that of a few existing profiling tools, using real and simulated sequence datasets. This report describes our benchmarking procedure and results.

For the background behind the MetaQuant project and detailed discussions of our modeling approach, algorithm, implementation and testing, we refer the reader to our previous technical reports [1] [2] [3].

2 Programs selected for comparative benchmarking

2.1 Overview

We compared the performance of MetaQuant against that of two other programs: "GASiC" [4] and "kallisto" [5]. We had originally intended to benchmark a third program, "GRAMMy" [6], instead of kallisto. However, GRAMMy turned out to have severe scalability issues, similar to those we discovered with GASiC (as described below). Other groups report that it is very difficult to use, and delivers less accurate results than GASiC on simulated test datasets (L. Schaeffer, personal communication). The newly developed program kallisto appears capable of handling much more realistic metagenomic datasets; therefore, we felt it would be more interesting and relevant to compare MetaQuant against kallisto.

While they differ greatly in their approaches, most programs used for profiling microbial abundances try to solve a similar set of problems. In general, these programs compare sequence reads or read pairs against a reference database of microbial genome sequences, or rely on an external read alignment program such as Bowtie2 [7] to do so. This produces a list of genome sequences that are "compatible" with each read, i.e. that the read could potentially be derived from. Typically, a read is compatible with multiple sequences in the reference genome database. The profiling tool must resolve this ambiguity, usually by estimating a probability for each read to be derived from each compatible reference genome. After performing these estimates for each read in the dataset, the profiler adds the probabilities for each genome to produce an estimate for the total number of reads that can be assigned to that genome. The assigned read counts are then used together with the genome sizes to infer the relative genome abundances.

2.2 GASiC

GASiC (an abbreviation for Genome Abundance Similarity Correction) addresses the ambiguous read match problem by first generating a similarity matrix $\mathbf{A}$ of values a_{ij} for each pair of genomes i, j in the reference database. a_{ij} represents the probability that a read selected randomly from genome i can be aligned to genome j. The probabilities are estimated by running a read simulator such as Mason [8] to generate 10,000 simulated reads for each reference genome, using Bowtie2 to align the reads against each of the other reference genomes, and counting the fractions of reads that have significant alignments. The read simulator parameters are chosen to emulate the known biases and error characteristics of the sequencing platform used to generate the real data. Once the similarity matrix $\mathbf{A}$ has been generated for a given reference database and sequencing platform, it can be used to estimate the true abundances c_j for a sequence dataset by solving the linear model $\mathbf{r} = \mathbf{Ac}$ where the element r_i of $\mathbf{r}$ is the observed number of reads aligning to genome i.

Because computing the similarity matrix for a reference database with N genomes requires $\frac{N(N-1)}{2}$ Bowtie2 runs, GASiC is not scalable for realistic metagenomic analysis problems, in which a minimally useful reference database contains thousands of genomes. Even for the "Illumina 100" benchmark described below, with a 100 genome database, building the similarity matrix required over 2.5 hours of computation on a 12 core server. To process a more reasonable database with 6000 genomes would take more than a year on the same hardware. Therefore, we only benchmarked GASiC on the "Illumina 100" dataset.

2.3 kallisto

While kallisto was originally developed for gene expression analysis using RNA-Seq data, its underlying algorithm is equally applicable to metagenomic abundance analysis [9]. It achieves fast performance by matching reads to target genome sequences using a k-mer hash index, rather than requiring an initial read alignment step, as do MetaQuant and GASiC. The result is a list of genome sequences that are compatible with each read, which is then used to construct a likelihood function for the whole dataset, parameterized by the genome abundances. An expectation-maximization (EM) algorithm is then used to find a maximum likelihood set of abundances.

Before profiling a dataset with kallisto, a user must construct an index for the reference database. The indexing process breaks each genome sequence into its component k-mers (with k typically set to 31) and constructs a colored de Bruijn graph (cDBG). In the cDBG, nodes represent k-mers, edges connect overlapping k-mers, and colored paths through the graph represent target genome sequences. Although indexing is fast for small reference databases, the time and memory required grow rapidly with database size. Indexing a database containing 238 bacterial genomes with a total of 800 megabases of sequence took 37.5 minutes, with memory usage peaking at 38.5 GB. Our standard RefMicrobial_Complete database (7,600 genomes, 11 gigabases) caused the indexer to run out of memory on a machine with 256 GB of RAM. Although the current version cannot

handle realistic microbial reference databases, the developers of kallisto claim that this scalability issue will be addressed in the near future [L. Pachter, personal communication].

3 Performance metrics and terminology

We assessed the performance of MetaQuant, GASiC and kallisto using a set of metrics that are commonly used in the metagenomic profiling literature. It will be helpful to first define terminology for the estimates produced by metagenomic profilers.

Relative genome abundance (RGA) refers to the fraction of complete genomes in a sample belonging to a given strain, species or genus; it is analogous to the molar concentration of a chemical compound. A *fragment* is a short DNA molecule, one end of which gets sequenced to produce a read, or both ends of which are sequenced to make a read pair. The *total fragment count* for a target sequence, strain, species or genus is the total number of fragments that can be mapped to the given target, including those that map ambiguously to multiple targets. The *assigned fragment count* is an estimate for the number of fragments that actually derive from a given target; the sum of these estimates should equal the total number of mapped fragments. The *assigned read fraction* or *sampling probability* is the assigned fragment count divided by the total number of mapped fragments. It is proportional to the product of the RGA and the *effective length* of the target genome. The effective length is roughly equal to the actual length of the target sequence, adjusted for fragment length and sequencer bias to reflect the probability of sampling a fragment from the target.

It is important to recognize that RGA values for different strains or species cannot simply be added to produce the corresponding species or genus level RGAs; instead, one must sum the corresponding assigned fragment counts, and then compute the RGA from the summed counts and genome sizes.

When the true abundances are known, one can compute various measures of deviation in the abundance estimates. Two that are commonly used in the metagenomic profiling literature are the *relative root mean square error* (RRMSE) and the *average relative error* (AVGRE). If c_t and $\hat{c}_t$ are respectively the true and estimated abundances for target t, and N targets are profiled, then these error rates are defined as:

$$RRMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{\hat{c}_t - c_t}{c_t}\right)^2}$$

$$AVGRE = \frac{1}{N}\sum_{t=1}^{N}\frac{|\hat{c}_t - c_t|}{c_t}$$

For the benchmarks described below, we computed the error rates for the RGAs estimated at the genome (strain), species, genus and family levels.

4 Benchmark datasets and results

4.1 Illumina 100 simulated dataset

The "Illumina 100" dataset [10] was one of several simulated datasets originally developed to assess the quality of genome assemblies generated by various programs. It consists of 27 million pairs of 75 bp reads, sampled from 100 reference bacterial and archaeal genomes. Quality scores were obtained from real Illumina data sets, and sequencing errors were generated randomly with probabilities based on the quality scores. The source genome sequences, with a few exceptions, included all chromosomes and plasmids from a given strain. For several species, genomes of multiple strains were included to test the ability of the profiler to distinguish reads from closely related genomes. The relative genome abundances used for the simulation were fairly even, ranging from 0.86% to 2.23%, with most strains having RGAs close to 1%.

We analyzed this dataset with all three programs - GASiC, kallisto and MetaQuant - using a reference database containing only the genome sequences used to generate the dataset. Although this is a contrived analysis scenario - in real life one would most likely not know the genomes present beforehand - the choice of reference database was necessitated by the limitations of GASiC, as described in section 2.2. Results from the three analyses at the genome and species level are shown in Figures 1, 2 and 3; the error rates at each taxonomic level are summarized in Tables 1 and 2. In the plots, labeled points indicate genomes or species whose abundance estimates differ from the true values by more than 0.2%.

Level	RRMSE		
	GASiC	kallisto	MetaQuant
genome	0.1421	0.0571	0.2674
species	0.0447	0.0217	0.1455
genus	0.0386	0.0223	0.0224
family	0.0237	0.0171	0.0172

Table 1: Root mean square relative errors (RRMSE) by taxonomic level for GASiC, kallisto and MetaQuant on Illumina 100 dataset

For this dataset, both GASiC and kallisto outperform MetaQuant at genome and species levels, while MetaQuant and kallisto are essentially equivalent and better than GASiC at genus and family levels. This is not too surprising, because of the different goals underlying the algorithms of the three programs. GASiC and kallisto both use a maximum likelihood approach to find relatively

Level	AVGRE		
	GASiC	kallisto	MetaQuant
genome	0.0673	0.0198	0.0894
species	0.0294	0.0122	0.0369
genus	0.0257	0.0121	0.0122
family	0.0190	0.0095	0.0094

Table 2: Average relative errors (AVGRE) by taxonomic level for GASiC, kallisto and MetaQuant on Illumina 100 dataset

unbiased estimates of read counts for each reference genome. By contrast, MetaQuant's Bayesian approach seeks to explain the data with a minimal set of genomes. When reads map ambiguously to similar reference genomes, MetaQuant is biased toward solutions that assign the majority of reads to the genome with which the largest proportion of reads are compatible. This underlying bias may be realistic for actual metagenomic datasets, but is problematic for artificial datasets. Thus, in the plots for MetaQuant, one can see several cases where one strain of a species has a relative abundance estimate near zero, while another strain of the same species has its abundance overestimated. Similarly, at the species level, most of the reads belonging to *Mycobacterium bovis* were assigned to *M. tuberculosis*, which has a nearly identical genome. To a lesser degree, GASiC also had trouble estimating abundances for strains of the same species. kallisto only had difficulty with the most similar pair of genomes, for two substrains of *E. coli* strain K-12.

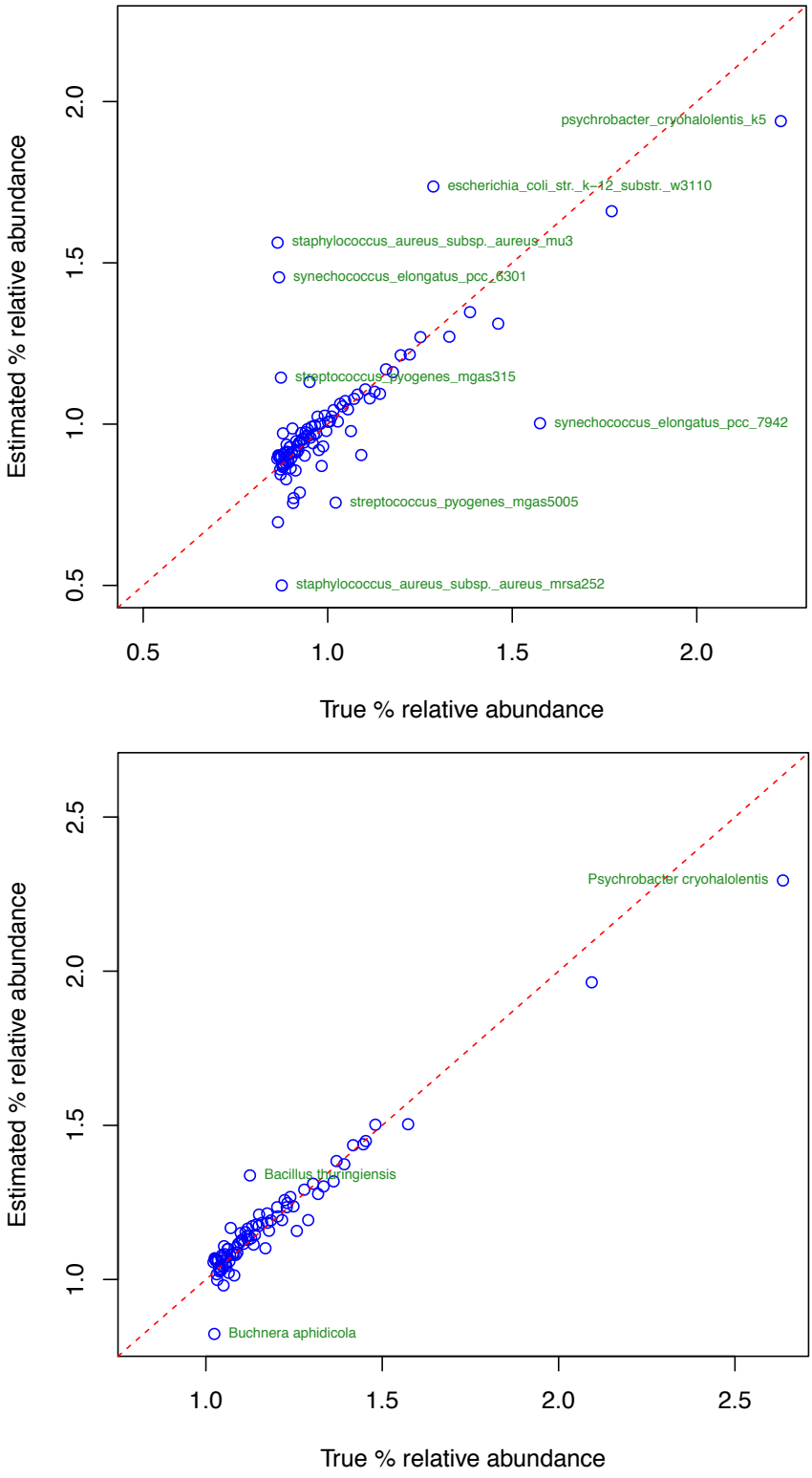

Figure 1: Relative abundances estimated by GASiC for each genome (top) or species (bottom), for the Illumina 100 simulated dataset, as a function of the input abundance

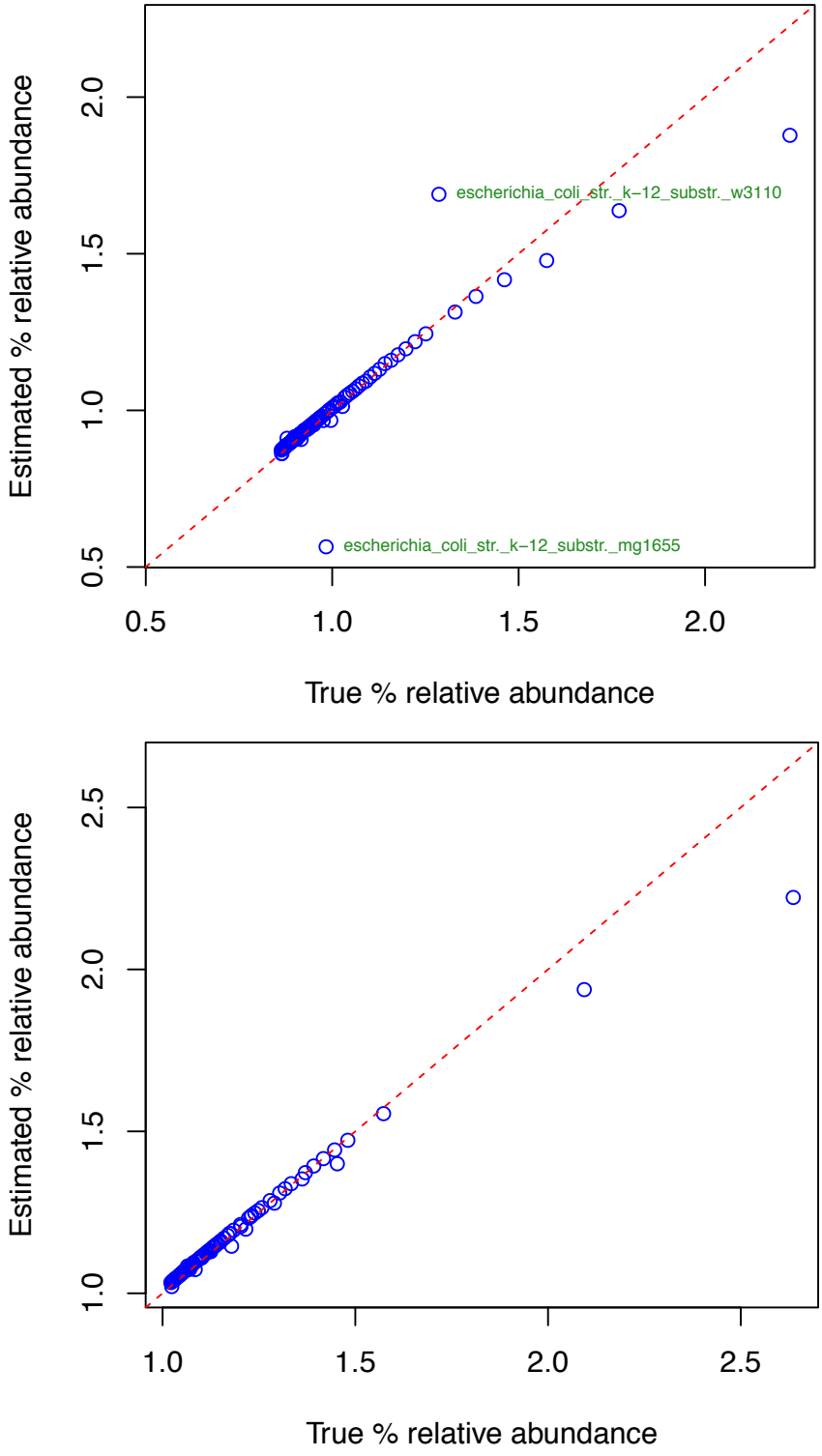

Figure 2: Relative abundances estimated by kallisto for each genome (top) or species (bottom), for the Illumina 100 simulated dataset, as a function of the input abundance

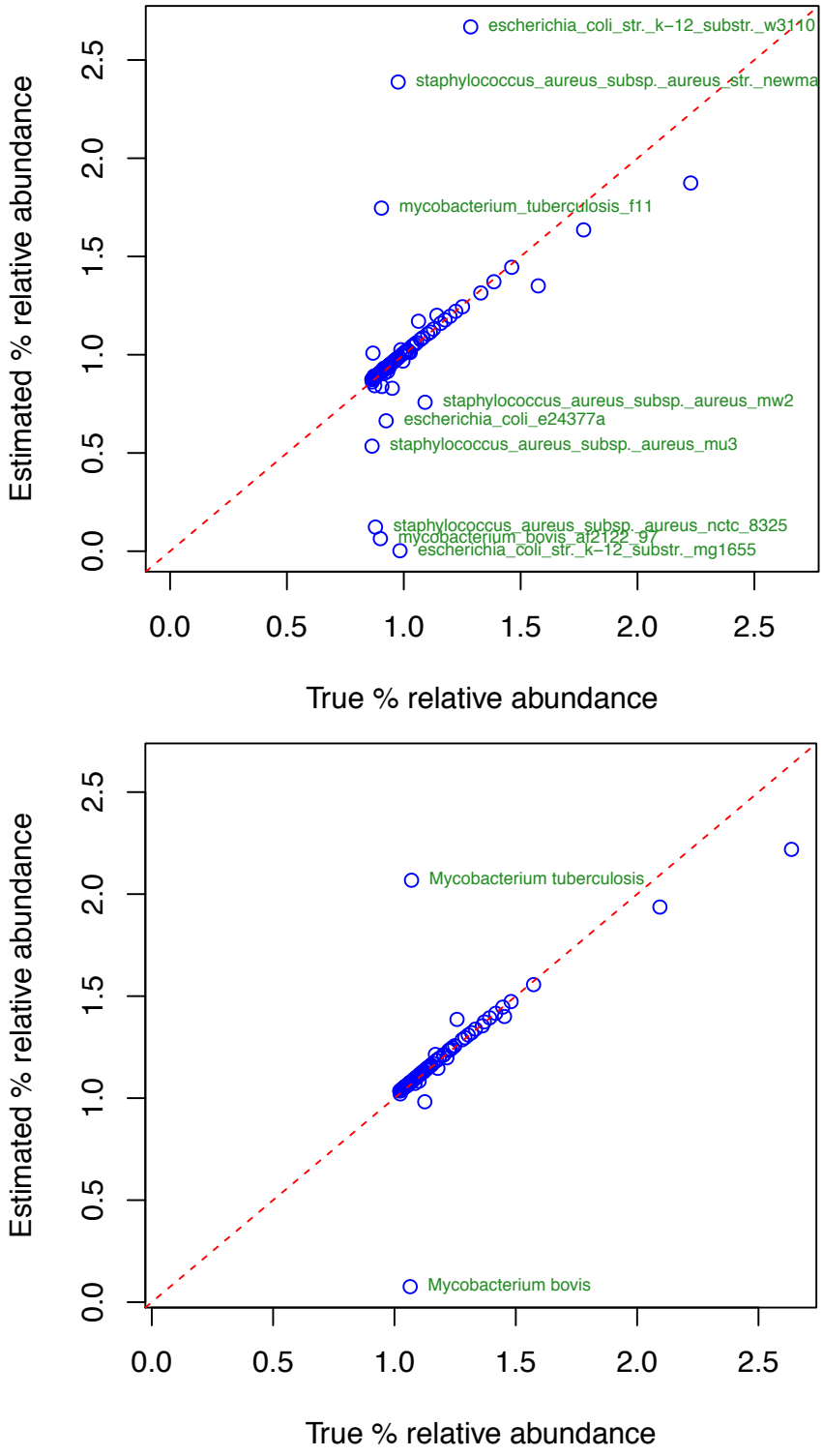

Figure 3: Relative abundances estimated by MetaQuant for each genome (top) or species (bottom), for the Illumina 100 simulated dataset, as a function of the input abundance

4.2 RefViral simulated dataset

The "RefViral" dataset was one of several simulated datasets developed for testing MetaQuant. It was generated by sampling 1,000,000 single-end 50-mer reads from randomly chosen viral genomes in the RefSeq database, following a MetaQuant model with parameters $\alpha = 1$ and $\beta = 10$. This produced a dataset containing sequences from 107 target genomes, with relative abundances falling off with rank approximately according to a power law with exponent -0.087, as shown in Figure 4. The relative abundances spanned about 6 orders of magnitude, with total sampled reads ranging from 1 to about 180,000. Therefore, this was a much more realistic simulation of a metagenomic dataset than the Illumina 100 data.

We ran MetaQuant and kallisto against this dataset, using the complete RefViral set of 5,301 genomes as the reference database. As we noted earlier, it would be infeasible to run GASiC with a database of this size, due to its need to align simulated reads from each genome against each other genome in the database. The results are shown in Figure 5. In both panels, the fitted relative abundances from each program are plotted against the input relative abundances from the simulation; the lower panel uses a log scale, to more clearly show the results for the low abundance genomes. For this dataset, MetaQuant and kallisto yielded nearly identical results, with excellent accuracy for most genomes. MetaQuant gave a slightly lower estimate for the second most abundant genome, assigning a small number of reads to another isolate of the same virus with 97% sequence identity to the correct isolate. kallisto underestimated abundances for several of the least abundant genomes, which were represented by only 1 to 4 reads in the dataset; and gave a zero estimate (not shown in the log scale plot) for one genome with 4 reads. In general, MetaQuant was more accurate than kallisto at the low end of the abundance range.

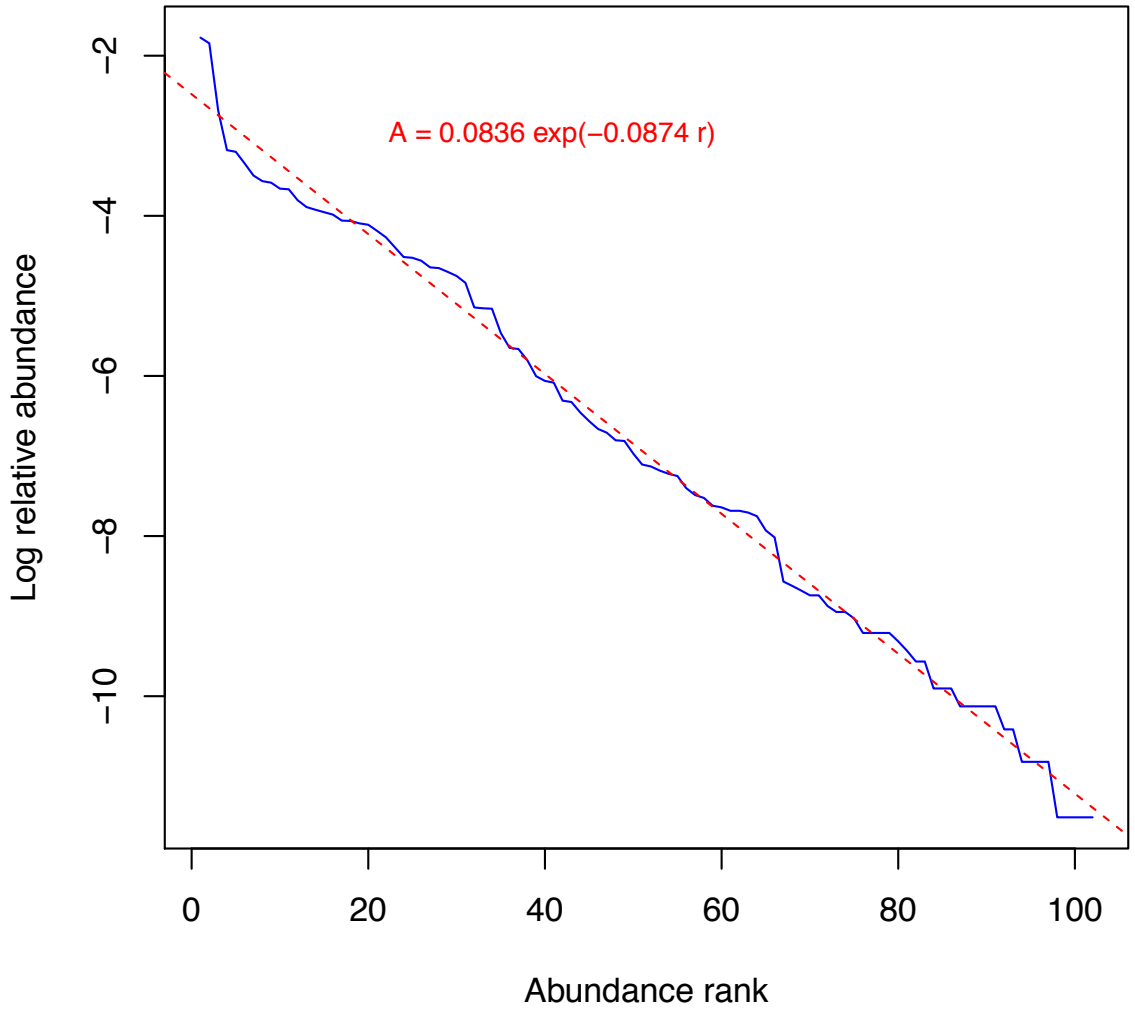

Figure 4: Log relative abundance vs rank for the source genome sequences used to generate the RefViral simulated dataset. The dashed line indicates the best power law fit to the abundance vs rank curve.

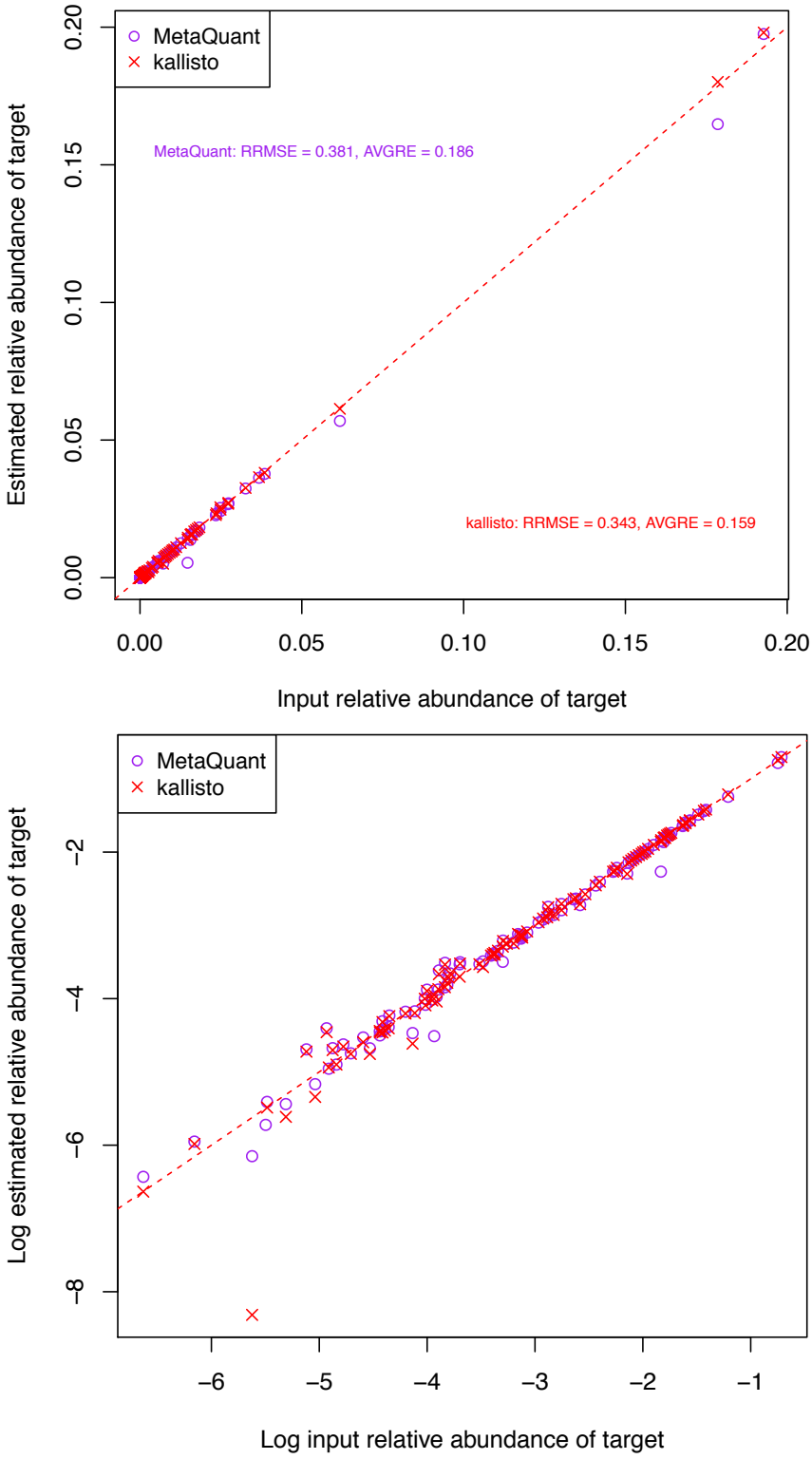

Figure 5: Relative abundances fit by MetaQuant and kallisto vs input values for the RefViral simulated dataset, on linear and $\log_{10}$ scales.

4.3 HMP staggered mock community dataset

The Human Microbiome Project (HMP) "mock community" datasets were developed by members of the HMP consortium to validate sequencing and analysis methods [11]. We used one of these (the "staggered" community dataset) for initial testing of MetaQuant. The dataset is described in detail in our testing report [3]; briefly, it was produced by sequencing a mixture of whole genome DNA from 20 bacterial species, one archaeal species and one fungus. The proportions of DNA in the "staggered" mixture were adjusted to provide different numbers of 16S gene copies per species, ranging from 1,000 to 1,000,000. Although the relative genome abundances in the mixture are known, there are known discrepancies in the representation of the various species in the sequence data, which are partly but not entirely explained by bias introduced in the library preparation step [12]. Nevertheless, this dataset is, to our knowledge, the best available example of real sequencing data derived from a large set of species covering a wide range of known abundances. The dataset is available on the Sequence Read Archive (accession SRR172903); it consists of 7.93 million single-end reads of length 75.

In order to benchmark MetaQuant against kallisto with this dataset, we had to choose a smaller set of reference sequences than the 11 gigabase "RefMicrobial-Complete" database used in our earlier testing, because kallisto is not currently able to index a database this large. We constructed a "MockPlus" reference set consisting of 238 genomes with the following taxonomic makeup:

- The 21 bacterial and archaeal strains known to be present in the mock community sample;
- Up to 4 additional strains, chosen at random, for each of the 21 species;
- Up to 4 other species in the same genus for each of the 18 genera represented in the mock community;
- 100 genomes from genera not represented in the mock community.

Our goal was to test the ability of kallisto and MetaQuant to discriminate between closely related strains and species. The number of genomes was limited for a few species and genera by the number of strains with available genome sequences. The resulting reference set contained a total of 789 megabases of sequence, which was well within the size that kallisto could handle.

The results from running kallisto and MetaQuant on the mock community dataset are shown in Figures 6, 7, 8 and 9. For each program, we compared the relative genome abundance estimates against the true abundances at the genome, species and genus level; the figures show the genome- and species-level results only. Table 3 shows the relative error rates computed at all three taxonomic levels. We found that kallisto and MetaQuant gave very similar results on this dataset. kallisto had slightly better error rates at the genome level, while MetaQuant did slightly better at the species level.

Level	RRMSE		AVGRE	
	kallisto	MetaQuant	kallisto	MetaQuant
genome	98.46	101.36	24.64	25.56
species	110.10	108.30	27.13	26.75
genus	6.45	6.55	1.95	1.93

Table 3: Root mean square and average relative errors by taxonomic level for kallisto and MetaQuant on HMP mock community dataset

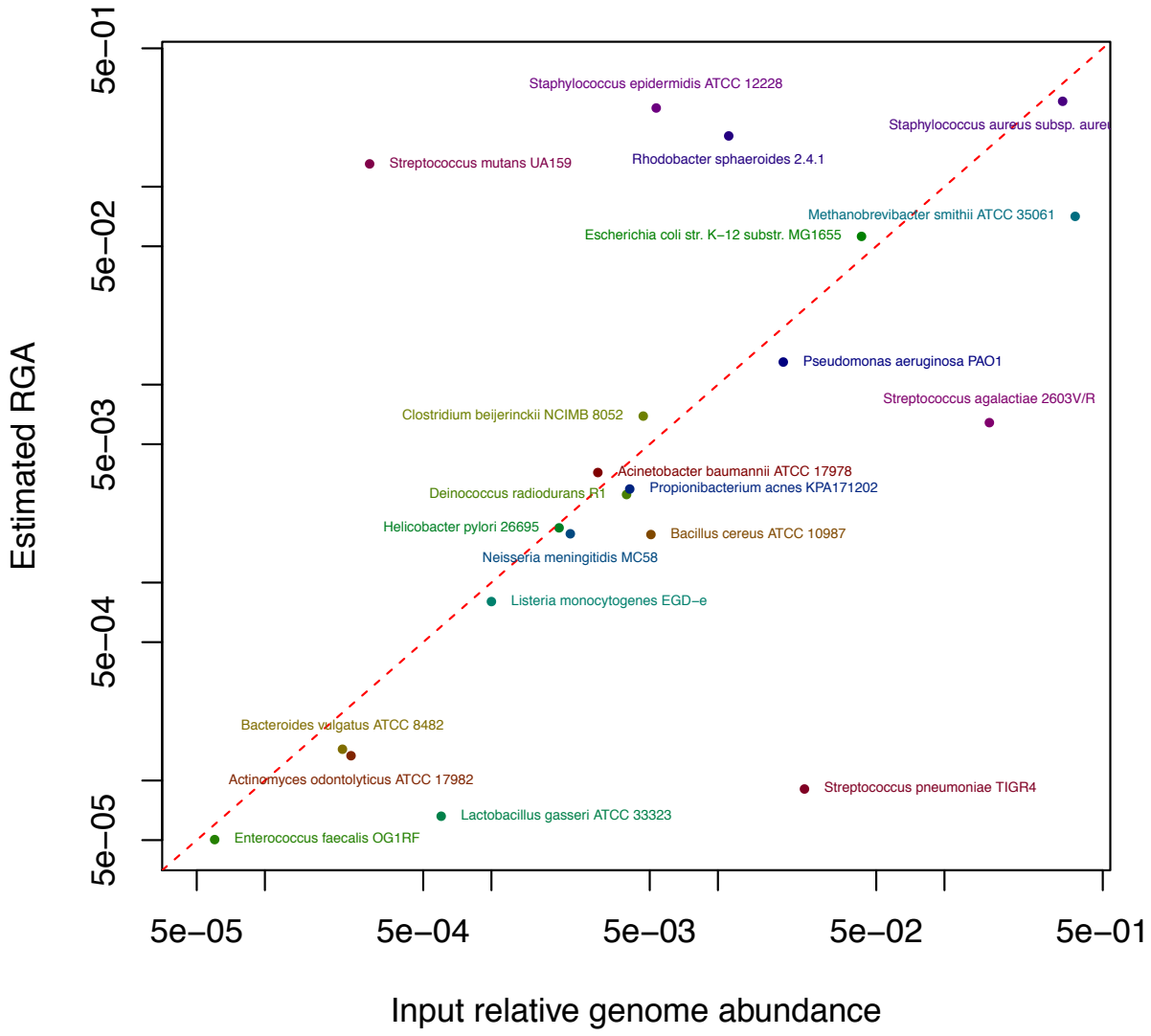

Figure 6: Relative abundances estimated by kallisto for each genome for the HMP staggered mock community dataset, as a function of the input abundance

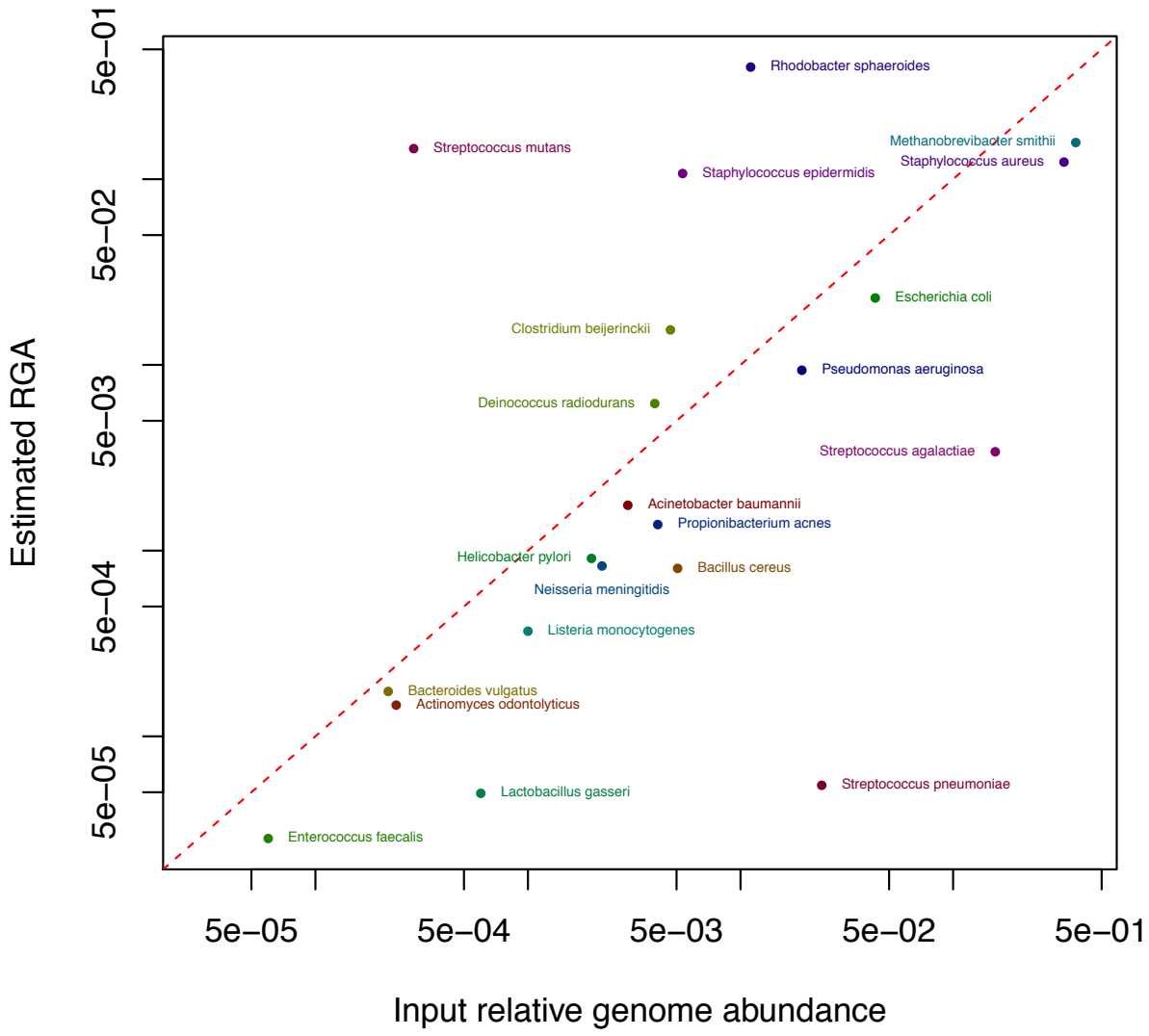

Figure 7: Relative abundances estimated by kallisto for each species for the HMP staggered mock community dataset, as a function of the input abundance

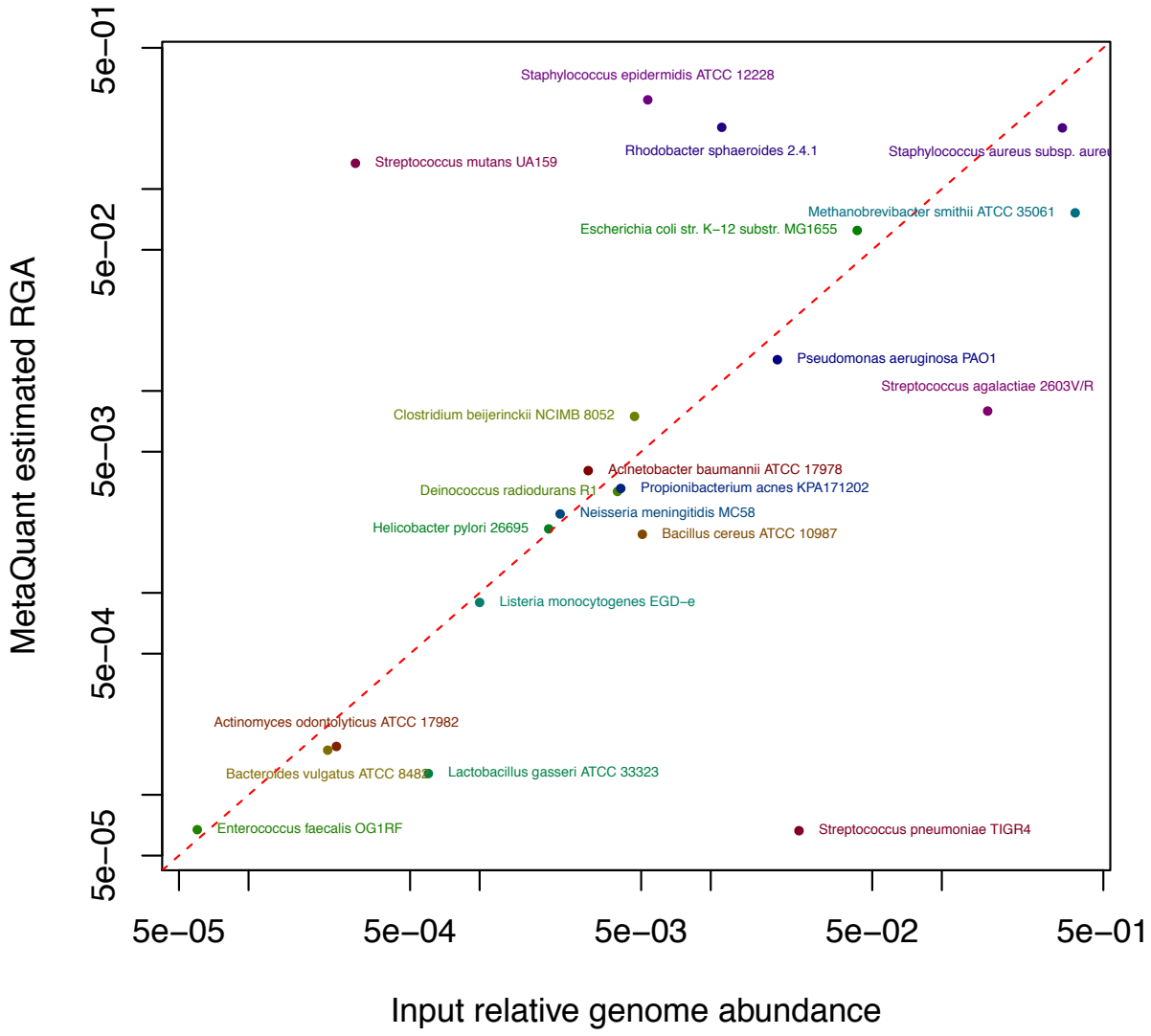

Figure 8: Relative abundances estimated by MetaQuant for each genome for the HMP staggered mock community dataset, as a function of the input abundance

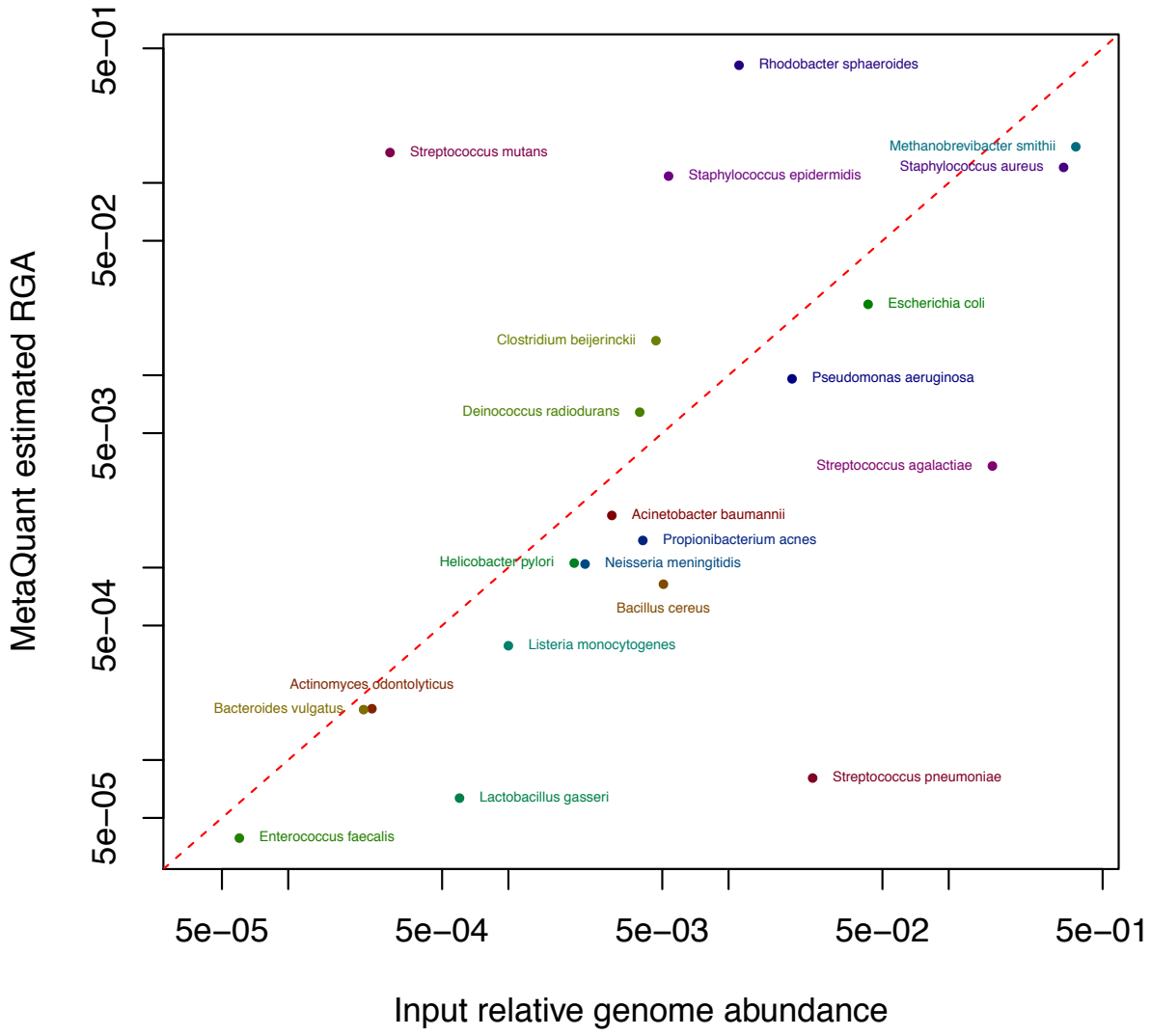

Figure 9: Relative abundances estimated by MetaQuant for each species for the HMP staggered mock community dataset, as a function of the input abundance

5 Discussion and conclusions

We draw two main conclusions from our benchmarking study. First, we found that many, if not all, open source metagenomic profiling tools described in the literature suffer from severe scalability issues. GASiC and GRAMMy in particular cannot be used with reference databases larger than a few hundred genome sequences, because the compute time and memory required to process the database goes up as the square of the database size. To be useful for analyzing samples of unknown composition, a reference database needs to include several thousand genomes. Therefore, while GASiC gave more accurate results than MetaQuant on a small simulated dataset, it was an artificial comparison because the reference database used contained the same genomes as were used to generate the simulated reads. It appears that GASiC and GRAMMy are only useful for analyzing samples in which the species present are known to be restricted to a small set.

The current version of kallisto also has a scalability problem, though it is not as severe as that of GASiC or GRAMMy. This is because the compute time required for kallisto to index a reference database is determined by the total number of sequence bases, rather than the number of genome sequences. Therefore, on a moderately large memory machine (256 GB RAM), it is able to index a database of all viral genome sequences from RefSeq (5,300 genomes with a total of 146 megabases), but not one containing all complete bacterial and archaeal genomes (7,600 genomes totalling 11 gigabases). On a larger server with 430 GB of RAM, it was possible to index a database of about 2,000 bacterial genomes with 4.3 gigabases of sequence (L. Schaeffer, personal communication). Useful metagenomic profiling analyses can be done with a database of this size, but it may require tailoring the database content to a target sample type, such as soil, indoor dust, human fecal samples or skin swabs.

As far as we are aware, MetaQuant is the only available metagenomic profiling tool that isn't restricted by these database size limitations. It is only subject to the constraints of the software used for sequence read alignment (i.e., Bowtie2), for which the indexing and search time is roughly linear with the database size. We have not run into problems building Bowtie2 indexes for the large microbial databases we have worked with, which range in size up to 27 gigabases.

The second lesson to be drawn from our benchmarking study is that the relative performance of metagenomic profiling tools depends strongly on the dataset used. One should be skeptical about performance results based on simulated datasets, because it is easy to tailor the dataset to emphasize the strengths of a particular tool and minimize its weaknesses. For example, MetaQuant gave less accurate results than either kallisto or GASiC on the Illumina 100 dataset, in which the component genomes had roughly equal abundances; but performed as well as or better than kallisto on the RefViral dataset, in which the genomes were present at a wide range of abundances. We believe the difference in accuracy is due in part to the Bayesian model underlying MetaQuant, which favors profiles in which the abundances span several orders of magnitude, and in which one strain dominates when a mixture of strains for one species is present. Since the Illumina 100 dataset draws equal numbers of reads from multiple genomes from each of several species, MetaQuant

gives skewed estimates for the genomes within those species.

The model underlying MetaQuant was designed to emulate the profiles of actual microbial communities, which tend to follow a power law abundance-vs-rank curve. The simulated RefViral dataset has such a profile, so the fact that MetaQuant performs well with it suggests that it should also do well with real metagenomic datasets. Unfortunately, there are few if any publically available WGS datasets whose abundance profiles have been well characterized by independent methods, such as qPCR. The best "real" datasets available, the HMP mock community datasets, contain *real* sequence data from *synthetic* microbial communities. The staggered mock community data provide a reasonable approximation to actual metagenomic data, in that the true abundances cover five orders of magnitude. This dataset has documented problems with library preparation and other biases. Nevertheless, the fact that MetaQuant and kallisto give similar results with this data, using completely orthogonal analysis approaches, shows that it is still a useful dataset for benchmarking. It would be of enormous benefit to metagenomics researchers if more well-characterized test datasets could be generated, from both real and synthetic microbial communities.

References

[1] Kevin McLoughlin. Modeling for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data. Lawrence Livermore National Laboratory Technical Report LLNL-TR-680776, Mar 2014.

[2] Kevin McLoughlin. Algorithm and Implementation for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data. Lawrence Livermore National Laboratory Technical Report LLNL-TR-680718, Jul 2014.

[3] Kevin McLoughlin. Software Implementation and Testing for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data. Lawrence Livermore National Laboratory Technical Report LLNL-TR-680721, Nov 2015.

[4] Martin S. Lindner and Bernhard Y. Renard. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Research*, 41(1):e10, 2012.

[5] Nicholas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal RNA-seq quantification. *Arxiv preprint arXiv:1505:02710*, 2015. http://arxiv.org/abs/1505.02710.

[6] Li C Xia, Jacob A Cram, Ting Chen, Jed A Fuhrman, and Fengzhu Sun. Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads. *PLoS ONE*, 6(12):e27992, 2011.

[7] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[8] M Holtgrewe. Mason – a read simulator for second-generation sequencing data. Technical Report TR-B-10-06, Institut fur Mathematik und Informatik, Freie Universitat Berlin, 2010.

[9] Lorian Schaeffer, Harold Pimentel, Nicholas L Bray, Pall Melsted, and Lior Pachter. Pseudoalignment for metagenomic read assignment. *Arxiv preprint arXiv:1510.07371*, 2015. http://arxiv.org/abs/1510.07371.

[10] Daniel Mende, Allison S Waller, Shinichi Sunagawa, Aino I Jarvelin, Michelle M Chan, Manimozhiyan Arumugam, Jeroen Raes, and Peer Bork. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE*, 7(2):e31386, 2012.

[11] Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PLoS ONE*, 7(6):e3931, 2012.

[12] Marcus B Jones, Sarah K Highlander, Ericka L Anderson, Weizhong Li, Mark Dayrit, Niels Klitgord, Martin M Fabani, Victor Seguritan, Jessica Green, David T Price, Shibu Yooseph, William Biggs, Karen E Nelson, and J Craig Venter. Library preparation methodology

can influence genomic and functional predictions in human microbiome research. *PNAS*, 112(45):14024–14029, 2015.